Une approche non paramétrique Bayesienne pour l'estimation de densité conditionnelle sur les rangs

Carine Hue*, Marc Boullé*

*France Télécom R & D; 2, avenue Pierre Marzin; 22307 Lannion cedex
Carine.Hue@orange-ftgroup.com; Marc.Boulle@orange-ftgroup.com

Résumé. Nous nous intéressons à l'estimation de la distribution des rangs d'une variable cible numérique conditionnellement à un ensemble de prédicteurs numériques. Pour cela, nous proposons une nouvelle approche non paramétrique Bayesienne pour effectuer une partition rectangulaire optimale de chaque couple (cible,prédicteur) uniquement à partir des rangs des individus. Nous montrons ensuite comment les effectifs de ces grilles nous permettent de construire un estimateur univarié de la densité conditionnelle sur les rangs et un estimateur multivarié utilisant l'hypothèse Bayesienne naïve. Ces estimateurs sont comparés aux meilleures méthodes évaluées lors d'un récent Challenge sur l'estimation d'une densité prédictive. Si l'estimateur Bayésien naïf utilisant l'ensemble des prédicteurs se révèle peu performant, l'estimateur univarié et l'estimateur combinant deux prédicteurs donne de très bons résultats malgré leur simplicité.

1 Introduction

Dans cette introduction, nous décrivons tout d'abord une situation particulière de l'apprentissage supervisé où l'on s'intéresse à prédire le rang d'une cible plutôt que sa valeur. Nous exposons ensuite deux approches qui permettent de passer d'une prédiction ponctuelle en régression à une description plus fine de la loi prédictive. Nous présentons ensuite notre contribution qui vise à fournir une estimation de la densité conditionnelle complète du rang d'une cible par une approche Bayesienne non paramétrique.

1.1 Régression de valeur et régression de rang

En apprentissage supervisé on distingue généralement deux grands problèmes : la classification supervisée lorsque la variable à prédire est symbolique et la régression lorsqu'elle prend des valeurs numériques. Dans certains domaines tels que la recherche d'informations, l'intérêt réside cependant plus dans le rang d'un individu par rapport à une variable plutôt que dans la valeur de cette variable. Par exemple, la problématique initiale des moteurs de recherche est de classer les pages associées à une requête et la valeur intrinsèque du score n'est qu'un outil pour produire ce classement. Indépendamment de la nature du problème à traiter, utiliser les rangs plutôt que les valeurs est une pratique classique pour rendre les modèles plus robustes aux valeurs atypiques et à l'hétéroscédasticité. En régression linéaire par exemple, un estimateur utilisant les rangs centrés dans l'équation des moindres carrés à minimiser est proposé

dans Hettmansperger et McKean (1998). L'apprentissage supervisé dédié aux variables ordinales est connu sous le terme de *régression ordinale* (cf Chou et Ghahramani (2005) pour un état de l'art). Dans la communauté statistique les approches utilisent généralement le modèle linéaire généralisé et notamment le modèle cumulatif (McCullagh, 1980) qui fait l'hypothèse d'une relation d'ordre stochastique sur l'espace des prédicteurs. En apprentissage automatique, plusieurs techniques employées en classification supervisée ou en régression métrique ont été appliquées à la régression ordinale : le principe de minimisation structurelle du risque dans Herbrich et al. (2000), un algorithme utilisant un perceptron appelé PRanking dans Crammer et Singer (2001) ou l'utilisation de machines à vecteurs de support dans Shashua et Levin (2002) Chu et Keerthi (2005). Les problèmes considérés par ces auteurs comprennent cependant une échelle de rangs fixée au préalable et relativement restreinte (de l'ordre de 5 ou 10). Autrement dit, le problème se ramène à prédire dans quel partile se trouve la cible, en ayant défini les partiles avant le processus d'apprentissage. On se rapproche alors plus d'un problème de classification et les algorithmes sont évalués selon leur taux de bonne classification ou sur l'erreur de prédiction entre le vrai partile et le partile prédit.

1.2 Vers une description plus complète d'une loi prédictive

Qu'il s'agisse de classification ou de régression, le prédicteur recherché est généralement ponctuel. On retient alors uniquement la classe majoritaire en classification ou l'espérance conditionnelle en régression métrique. Ces indicateurs peuvent se révèler insuffisants, notamment pour prédire des intervalles de confiance mais également pour la prédiction de valeurs extrêmes. Dans ce contexte, la régression quantile ou l'estimation de densité permettent de décrire plus finement la loi prédictive.

La régression quantile vise à estimer plusieurs quantiles de la loi conditionnelle. Pour α réél dans $[0, 1]$, le quantile conditionnel $q_\alpha(x)$ est défini comme le réél le plus petit tel que la fonction de répartition conditionnelle soit supérieure à α. Reformulé comme la minimisation d'une fonction de coût adéquate, l'estimation des quantiles peut par exemple être obtenue par l'utilisation de splines (Koenker, 2005) ou de fonctions à noyaux (Takeuchi et al., 2006). Les travaux proposés dans Chaudhuri et al. (1994); Chaudhuri et Loh (2002) combinent un partitionnement de l'espace des prédicteurs selon un arbre et une approche polynômiale locale. La technique récente des forêts aléatoires est étendue à l'estimation des quantiles conditionnels dans Meinshausen (2006). En régression quantile, les quantiles que l'on souhaite estimer sont fixés à l'avance et les performances sont évaluées pour chaque quantile.

Les techniques d'estimation de densité visent à fournir un estimateur de la densité conditionnelle $p(y|x)$. L'approche paramétrique présuppose l'appartenance de la loi conditionnelle à une famille de densités fixée à l'avance et ramène l'estimation de la loi à l'estimation des paramètres de la densité choisie. Les approches non paramétriques, qui s'affranchissent de cette hypothèse, utilisent généralement deux principes : d'une part, l'estimateur de la densité est obtenu en chaque point en utilisant les données contenues dans un *voisinage* autour de ce point ; d'autre part, une hypothèse est émise sur la forme recherchée localement pour cet estimateur. Très répandues, les méthodes dites à noyau définissent le voisinage de chaque point en convoluant la loi empirique des données par une densité à noyau centrée en ce point. La forme du noyau et la largeur de la fenêtre sont des paramètres à régler. Une fois la notion de voisinage définie, les techniques diffèrent selon la famille d'estimateurs visée : l'approche polynômiale locale (Fan et al., 1996) regroupe les estimateurs constants, linéaires ou d'ordre supérieur. On

peut également chercher à approximer la densité par une base de fonctions splines. Cette démarche d'estimation de la loi complète a déjà été adoptée en régression ordinale dans Chu et Keerthi (2005) en utilisant des processus Gaussiens dans un cadre Bayésien .

1.3 Notre contribution

Nous proposons ici une approche Bayésienne non paramétrique pour l'estimation de la loi conditionnelle du rang d'une cible numérique. Notre approche utilise la statistique d'ordre en amont du processus d'apprentissage. La manipulation exclusive des rangs au détriment des valeurs rend notre estimateur invariant par toute transformation monotone des données et peu sensible aux valeurs atypiques. Si le problème étudié ne s'intéresse pas à des variables ordinales mais à des variables numériques on peut bien entendu s'y ramener en calculant les rangs des exemples à partir de leurs valeurs. Contrairement aux problèmes habituellement traités en régression ordinale, on considère en amont de l'apprentissage l'échelle globale des rangs de 1 au nombre d'exemples dans la base. Notre méthode effectue un partitionnement 2D optimal qui utilise l'information d'un prédicteur pour mettre en évidence des plages de rangs de la cible dont le nombre n'est pas fixé à l'avance. Disposant d'un échantillon de données de taille finie N et ne souhaitant pas émettre d'hypothèse supplémentaire sur la forme de la densité prédictive, nous nous restreignons à des densités conditionnelles sur les rangs constantes sur chaque cellule. Notre estimateur se ramène donc à un vecteur d'estimateurs de quantiles de la loi conditionnelle sur les rangs. A la différence de la régression quantile, le choix des quantiles n'est pas décidé au préalable mais est guidé par les partitions obtenues.

Suite à ce positionnement, la seconde partie est consacrée à la description de l'approche MODL pour le partitionnement 1D en classification supervisée puis pour le partitionnement 2D en régression. Nous détaillons dans la troisième partie comment obtenir un estimateur univarié et un estimateur multivarié Bayesien naïf de la densité prédictive à partir des effectifs de ces partitionnements. Dans la quatrième partie, les estimateurs obtenus sont testés sur quatre jeux de données proposés lors d'un challenge récent et sont comparés avec les autres méthodes en compétition. La dernière partie est consacrée à la conclusion.

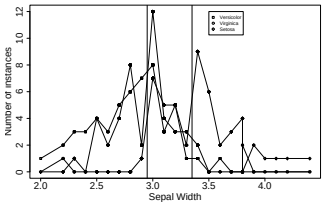
2 La méthode de partitionnement 2D MODL pour la régression

Nous présentons ici l'approche MODL pour le partitionnement 1D en classification supervisée (Boullé, 2006) puis son extension pour le partitionnement 2D en régression.

2.1 L'approche MODL pour la classification supervisée

En classification supervisée, l'objectif d'une méthode de partitionnement est de discrétiser le domaine d'un prédicteur à valeurs numériques, de manière à mettre en valeur le maximum d'informations sur la variable cible sur chaque intervalle. Un compromis doit être trouvé entre la qualité de l'information prédictive, i.e. la capacité de la discrétisation à discriminer les classes cibles, et la qualité statistique i.e. la robustesse de la discrétisation en généralisation.
Par exemple, le nombre d'instances de chaque classe de la base de données Iris (D.J. Newman

et Merz, 1998) est tracé en fonction de la largeur des sépales, à gauche de la Fig. 1. La discrétisation consiste à trouver la partition de $[2.0, 4.4]$ qui donne le maximum d'informations sur la répartition des trois classes connaissant l'intervalle de discrétisation.



	]2.0,2.95[	[2.95, 3.35[	[3.35, 4.4[
Versicolor	34	15	1
Virginica	21	24	5
Setosa	2	18	30
Total	57	57	36

FIG. 1 – *Discrétisation MODL de la variable Largeur de sépale pour la classification du jeu de données Iris en trois classes.*

L'approche MODL (Boullé, 2006) considère la discrétisation comme un problème de sélection de modèle. Ainsi, une discrétisation est considérée comme un modèle paramétré par le nombre d'intervalles, leurs bornes et les effectifs des classes cible sur chaque intervalle. La famille de modèles considérée est l'ensemble des discrétisations possibles. On dote cette famille d'une distribution *a priori* hiérarchique et uniforme à chaque niveau selon laquelle :
- le nombre d'intervalles de la discrétisation est distribuée de manière équiprobable entre 1 et le nombre d'exemples ;
- étant donné un nombre d'intervalles, les distributions des effectifs sur chaque intervalle sont équiprobables ;
- étant donné un intervalle, les distributions des effectifs par classe cible sont équiprobables ;
- les distributions des effectifs par classe cible pour chaque intervalle sont indépendantes les unes des autres.

Adoptant une approche Bayesienne, on recherche alors le modèle le plus vraisemblable connaissant les données. En utilisant la formule de Bayes et le fait que la probabilité du jeu de données soit constante quel que soit le modèle, le modèle visé est celui qui maximise le produit $p(\text{modèle}) \times p(\text{données}|\text{modèle})$.
Formellement, on note N le nombre d'exemples, J le nombre de classes cible, I le nombre d'intervalles d'une discrétisation, $N_{i.}$ le nombre d'exemples dans l'intervalle i et N_{ij} le nombre d'exemples dans l'intervalle i appartenant à la $j$ème classe cible. En classification supervisée, les nombres d'exemples N et de classe cible J sont connus. On caractérise donc un modèle de discrétisation par les paramètres $\left\{I, \{N_{i.}\}_{1 \leq i \leq I}, \{N_{ij}\}_{1 \leq i \leq I, 1 \leq j \leq J}\right\}$. On remarque que, dans la mesure où une discrétisation est caractérisée par les effectifs des intervalles, un tel modèle est invariant par toute transformation monotone des données.
En considérant l'ensemble des discrétisations possibles sur lequel on adopte la distribution *a priori* hiérarchique uniforme décrite précédemment, on obtient que le log négatif du produit $p(\text{modèle}) \times p(\text{données}|\text{modèle})$ peut s'écrire sous la forme du critère d'évaluation suivant (1) :

$$\log N + \log \left(\begin{array}{c} N+I-1 \\ I-1 \end{array} \right) + \sum_{i=1}^{I} \log \left(\begin{array}{c} N_{i.}+J-1 \\ J-1 \end{array} \right) + \sum_{i=1}^{I} \log \frac{N_i!}{N_{i1}!N_{i2}!\ldots N_{iJ}!} \quad (1)$$

Les trois premiers termes évaluent le log négatif de la probabilité *a priori* : le premier terme correspond au choix du nombre d'intervalles, le second au choix de leurs bornes et le troisième terme décrit la répartition des classes cibles sur chaque intervalle. Le dernier terme est le log négatif de la vraisemblance des données conditionellement au modèle. On trouvera tous les détails nécessaires à l'obtention de ce critère dans Boullé (2006).
Pour pouvoir traiter d'importants volumes de données, une heuristique gloutonne ascendante est proposée dans Boullé (2006). Afin d'améliorer la solution obtenue, des post-optimisations sont menées au voisinage de cette solution, en tentant des enchaînements de coupures et de fusion d'intervalles. L'additivité du critère permet de réduire la complexité de l'algorithme en $O(JNlog(N))$, post-optimisations comprises.
Pour l'exemple donné à gauche de la Fig. 1, la partition optimale obtenue est décrite par la table de contingence figurant à droite. Les bornes des intervalles sont obtenues en moyennant les valeurs du dernier individu d'un intervalle et du premier individu de l'intervalle suivant. On peut en déduire des règles telles que "'pour une largeur de sépale comprise entre 2.0 et 2.95, la probabilité d'occurence de la classe Versicolor est de $34/57 = 0.60$".

2.2 L'approche MODL pour la régression

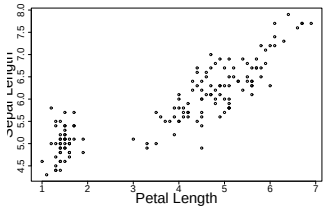


FIG. 2 – *Diagramme de dispersion de la variable Longueur de sépale en fonction de Longueur de pétale pour le jeu de données Iris.*

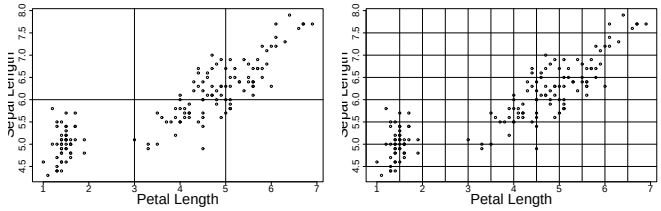


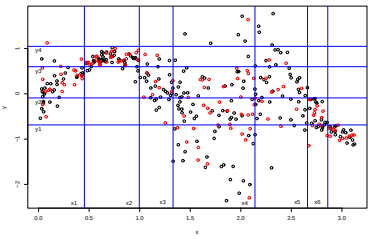
FIG. 3 – *Deux grilles de discrétisation à* 6 *et* 96 *cellules mettant en valeur la corrélation entre les variables Longueur de pétale et Longueur de sépale du jeu de données Iris.*

Pour illustrer le problème du partitionnement 2D lorsque le prédicteur et la cible sont des variables numériques, nous présentons en Fig. 2 le diagramme de dispersion des variables longueur de pétale et longueur de sépale du jeu de données Iris. La figure montre que les iris dont la longueur de pétale est inférieure à deux cm ont toujours des sépales de longueur inférieure

à six cm. Si l'on sépare les valeurs de la variable cible longueur de sépale en deux intervalles (inférieur et supérieur à six cm), on peut décrire la loi de répartition de cette variable conditionnellement au prédicteur longueur de pétale à l'aide d'une discrétisation du prédicteur comme en classification supervisée.
L'objectif d'une méthode de partitionnement 2D est de décrire la distribution des rangs d'une variable cible numérique étant donné le rang d'un prédicteur. La discrétisation du prédicteur et de la cible comme illustrée en Fig. 3 permettent une telle description. Par rapport au cas de la classification supervisée, une partition 2D (ou grille) est modélisée par un paramètre supplémentaire, le nombre d'intervalles de la variable cible. Un compromis doit être trouvé entre la qualité de l'information de corrélation détectée entre les deux variables et la capacité de généralisation de la grille. Formellement, on caractérise un modèle de discrétisation 2D par les paramètres $\left\{I, J, \{N_{i.}\}_{1\leq i\leq I}, \{N_{ij}\}_{1\leq i\leq I; 1\leq j\leq J}\right\}$. Le nombre d'exemples dans chaque intervalle cible est noté $N_{.j}$ et se déduit par sommation des nombres d'exemples des cellules d'un même intervalle de rang. On adopte l'*a priori* suivant pour ces paramètres :
1. les nombres d'intervalles I et J sont indépendents et distribués uniformément entre 1 et N ;
2. pour un nombre d'intervalles I donné, toutes les partitions en intervalles des rangs du prédicteur sont équiprobables ;
3. pour un intervalle source donné, toutes les distributions des exemples sur les intervalles cibles sont équiprobables ;
4. les distributions des exemples sur les intervalles cible pour chaque intervalle du prédicteur sont indépendantes les unes des autres ;
5. pour un intervalle cible donné, toutes les distributions des rangs sont équiprobables.
En utilisant le modèle de discrétisation 2D et la loi *a priori* définis précedemment, on peut écrire le logarithme négatif du produit $p(\mathrm{M}) \times p(\text{données}|\mathrm{M})$ sous la forme du critère (2) pour un modèle de discrétisation M :

$$\begin{aligned} c_{reg}(M) = 2\log(N) + \log\left(\begin{array}{c} N+I-1 \\ I-1 \end{array}\right) + \sum_{i=1}^{I} \log\left(\begin{array}{c} N_i+J-1 \\ J-1 \end{array}\right) \\ + \sum_{i=1}^{I} \log \frac{N_i!}{N_{i,1}!N_{i,2}!\ldots N_{i,J}!} + \sum_{j=1}^{J} \log N_{.j}! \end{aligned} \tag{2}$$

Par rapport au critère obtenu dans le cas de la classification supervisée, il y a un terme supplémentaire égal à $\log(N)$ pour la prise en compte du nombre d'intervalles cible selon la loi *a priori* et un terme additif en $\log(N_{.j}!)$ qui évalue la vraisemblance de la distribution des rangs des exemples dans chaque intervalle cible.
Nous présentons succintement l'algorithme adopté pour minimiser ce critère. Nous débutons avec une partition initiale aléatoire de la cible puis, tant que le critère décroît, nous optimisons alternativement la partition 1D du prédicteur pour la partition fixée de la cible et la partition 1D de la cible pour la partition fixée du prédicteur. Nous répétons le processus pour plusieurs partitions initiale aléatoire de la cible et nous retournons la partition qui minimise le critère. En pratique, la converge s'effectue très rapidement, en deux ou trois itérations. Les discrétisations 1D sont effectuées selon l'algorithme d'optimisation utilisé pour la classification supervisée.
La valeur du critère (2) pour un modèle de discrétisation donné est relié à la probabilité que ce modèle explique la cible. C'est donc un bon indicateur pour évaluer les prédicteurs dans un problème de régression. Les prédicteurs peuvent être classés par probabilité décroissante de



	P1	P2	P3	P4	P5	P6	P7	Tot.
C5	1	0	0	6	5	0	0	12
C4	4	68	5	2	9	0	0	88
C3	38	5	26	32	24	0	0	125
C2	9	0	6	33	15	21	0	84
C1	0	0	0	34	4	9	28	75
Total	52	73	37	107	57	30	28	384

FIG. 4 – *Diagramme de dispersion, partitionnement 2D et effectifs de la grille MODL pour le jeu de données Synthetic*

leur capacité à expliquer la cible (Boullé et Hue, 2006). Afin de fournir un indicateur normalisé, on considère la fonction suivante du critère c_{reg} : $g\left(M\right) = 1 - \frac{c_{reg}(M)}{c_{reg}(M_\emptyset)}$, où $M_\emptyset$ est le modèle vide avec un unique intervalle prédicteur et cible. Le taux de compression $g(M)$ prend ses valeurs entre 0 and 1, dans la mesure où le modèle vide est toujours évalué dans notre algorithme d'optimisation. Le taux vaut 0 pour le modèle vide et est maximal pour la meilleure description des rangs cible conditionnellement au prédicteur.

3 De la discrétisation 2D à l'estimation de densité conditionnelle sur les rangs

3.1 Cas univarié

Le passage du partitionnement 2D à l'estimation de densité conditionnelle univariée est illustrée sur un jeu de données synthétique proposé lors du récent Challenge *Predictive Uncertainty in Environmental Modelling* (Cawley et al., 2006) qui comporte $N = 384$ exemples et un seul prédicteur. Le diagramme de dispersion ainsi que la partition MODL obtenue sont représentés en Fig. 4. Les intervalles de rangs [1] sont notés P_i pour $i = 1, \ldots, 7$ pour le prédicteur et $C_j, j = 1, \ldots, 5$ pour la cible. Comme pour le partitionnement 1D, les bornes des intervalles notées $x_1, \ldots, x_6$ et $y_1, \ldots, y_4$ sont obtenues par moyennage des valeurs du dernier individu d'un intervalle et du premier individu de l'intervalle suivant. Soit x la valeur du prédicteur pour un nouvel individu et P_i^x la plage de rangs à laquelle appartient le rang de x. Les effectifs de la grille nous permettent de calculer directement la probabilité que le rang de la cible de ce nouvel individu soit dans un intervalle donné C_j :

$$P_{Modl}\left(rg(y) \in C_j | rg(x) \in P_i^x\right) = \frac{N_{ij}}{N_{i.}}. \tag{3}$$

En supposant la densité conditionnelle sur les rangs constante sur chaque plage de rangs cible que délimite la grille, on obtient une expression pour les probabilités élémentaires

$$P_{Modl}(k \leq rg(y) < k+1 | rg(x) \in P_i^x) = \frac{N_{ij}}{N_{i.}N_{.j}} \tag{4}$$

[1] fermés à gauche et ouverts à droite

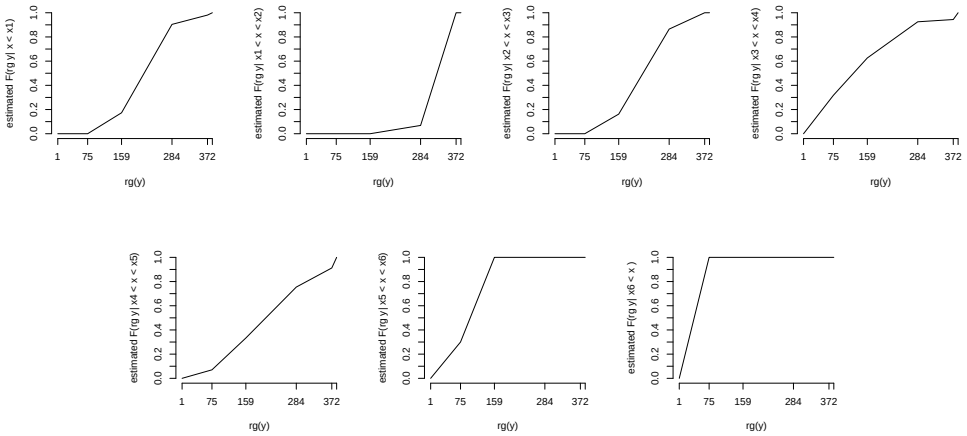


FIG. 5 – *Estimations MODL de la fonction de répartition conditionnelle univariée sur les rangs pour les sept intervalles de discrétisation du prédicteur*

où k est dans le $j$ème intervalle de rang C_j ($\sum_{l=1}^{j-1} N_{.l} \leq k \leq \sum_{l=1}^{j} N_{.l} - 1$). On obtient un estimateur de la fonction de répartition conditionnelle sur les rangs en cumulant ces probabilités élémentaires :

$$P_{Modl}(rg(y) < k | rg(x) \in P_i^x) = \sum_{l=1}^{j-1} \frac{N_{il}}{N_{i.}} + \left(k - \sum_{l=1}^{j-1} N_{.l} \right) \times \frac{N_{ij}}{N_{i.} N_{.j}} \tag{5}$$

où $\sum_{l=1}^{j-1} N_{.l} \leq k < \sum_{l=1}^{j} N_{.l}$.

Les estimateurs MODL de la fonction de répartition conditionnelle au prédicteur sont tracés pour chacun des sept intervalles du prédicteur en Fig 5.

3.2 Cas multivarié

Dans le cas de P ($P > 1$) prédicteurs, on peut en première approche construire un estimateur sous l'hypothèse Bayesienne naïve que les prédicteurs soient indépendants conditionnellement à la cible. Soient $(x_1, \ldots, x_P)$ les coordonnées d'un nouvel individu dans l'espace des prédicteurs et P_i^x l'intervalle de discrétisation auquel appartient chaque composante x_i. Sous l'hypothèse Bayesienne naïve, la probabilité élémentaire multivarié s'écrit alors :

$$\begin{aligned} &P(k \leq rg(y) < k+1 | (rg(x_1), \ldots, rg(x_P)) \in (P_1^x, \ldots, P_P^x)) \\ &\quad \propto P(k \leq rg(y) < k+1) \prod_{l=1}^{P} P(rg(x_l) \in P_l^x | k \leq rg(y) < k+1) \\ &\quad = P(k \leq rg(y) < k+1) \prod_{l=1}^{P} \frac{P(k \leq rg(y) < k+1 | rg(x_l) \in P_l^x) P(rg(x_l) \in P_l^x)}{P(k \leq rg(y) < k+1)} \end{aligned} \tag{6}$$

Cette dernière expression peut être estimée grâce aux effectifs des partitionnements 2D. On peut en effet directement estimer le premier facteur $P(k \leq rg(y) < k+1)$ par la probabilité empirique $1/N$. En ce qui concerne le produit, le premier facteur du numérateur s'obtient selon le même principe que les probabilités univariées en (4) en considérant la plage de rangs à laquelle appartient y après fusion des partitions de la cible induites par chaque prédicteur. Soit $\mathcal{J}$ le nombre d'intervalles cible pour la partition cible résultant de cette fusion et $N_{.j}^{M}$ l'effectif de chaque plage de rangs C_j^M pour $j = 1, \ldots, \mathcal{J}$. Par construction, le partitionnement de la cible associé à chaque prédicteur est inclus dans ce partitionnement "multivarié" et l'on note N_{ij}^{lM} l'effectif de la cellule associée au $i$ème intervalle du $l$ème prédicteur et au $j$ème intervalle cible du partitionnement multivarié. Chaque fraction du produit précédent peut alors s'estimer par $\frac{\frac{N_{ij}^{lM}}{N_{i.}^{l} N_{.j}^{M}} \frac{N_{i.}^{l}}{N}}{N} = \frac{N_{ij}^{lM}}{N_{.j}^{M}}$.

3.3 Evaluation d'un estimateur de la densité conditionnelle sur les rangs

En apprentissage supervisé, les fonctions de score les plus couramment utilisés pour évaluer un prédicteur sont le score logarithmique et le score quadratique qui prennent différentes formes selon la tâche visée (classification, régression métrique ou ordinale) et l'approche adoptée (déterministe ou probabiliste). En régression ordinale, les approches déterministes citées en introduction sont évaluées sur l'erreur quadratique moyenne entre le rang prédit et le rang vrai en considérant les rangs comme des entiers consécutifs. Pour une prédiction probabiliste, le score logarithmique du NLPD est également utilisé dans Chu et Keerthi (2005) où

$$NLPD(\hat{p}, D) = \frac{1}{N} \sum_{l=1}^{L} -\log(\hat{p}(y_l|x_l)) \tag{7}$$

pour le jeu de données $D = (x_l, y_l)_{l=1,\ldots,L}$. Nous utilisons cette fonction de score pour notre estimateur de la densité conditionnelle sur les rangs. Pour un nouvel individu (x_l, y_l), on calcule son rang d'insertion k dans l'échantillon d'apprentissage et on estime $\hat{p}(rg(y)|rg(x))$ par la probabilité élémentaire $\hat{P}_{Modl}(k \leq rg(y) < k+1|rg(x))$ décrite en (5) et en (6) .

4 Evaluation expérimentale

Comme exposé en introduction, notre approche se distingue des problèmes habituellement traités en régression ordinale du fait qu'on estime la distribution pour l'échelle globale des rangs et non pour une plage de rangs restreinte à 5 ou 10 rangs distincts. D'autre part, peu de méthodes fournissent un estimateur de la loi complète. Afin de positionner la méthode, nous avons donc choisi de la comparer en premier lieu à d'autres estimateurs de densité conditionnelle sur les valeurs. Nous avons pour cela choisi les quatre jeux de données proposés lors du récent Challenge Predictive Uncertainty in Environmental Modelling organisé en 2006 [2] et décrits dans le tableau 1. Notre approche étant par nature régularisée et n'ayant aucun paramètre de réglage , nous prenons le parti d'utiliser les jeux de données d'apprentissage et de validation

[2] http ://theoval.cmp.uea.ac.uk/ gcc/competition/

	Synthetic	SO2	Precip	Temp
Nbre de prédicteurs	1	27	106	106
Nbre d'ex en apprentissage	384	22956	10546	10675
Nbre d'ex en test	1024	7652	3517	3560

TAB. 1 – *Jeux de données du Challenge Predictive Uncertainty in Environmental Modelling avec le nombre de prédicteurs, le nombre d'individus utilisés en apprentissage et en test.*

pour calculer les partitionnements 2D optimaux. Outre les effectifs de chaque grille, ces partitionnements fournissent également un indice de compression défini en (2.2) qui nous permet d'évaluer de manière univariée chaque prédicteur puis de les classer par importance prédictive. Pour les jeux de données réels nous avons calculé l'estimateur multivarié Bayesien naïf utilisant l'ensemble des prédicteurs mais également l'estimateur univarié utilisant uniquement le prédicteur dont l'importance prédictive MODL était la plus élevée ainsi que l'estimateur Bayesien naïf utilisant le couple de variables le plus informatif.
Connaissant les valeurs associées aux rangs, chaque estimateur sur les rangs nous fournit un estimateur de la fonction de répartition conditionnelle en les N valeurs cibles de l'échantillon. Si l'on note $y_{(k)}$ la valeur cible de l'individu de rang k, on a en effet :

$$P_{Modl}(y < y_{(k)}|x) = P_{Modl}(rg(y) < k|x) \tag{8}$$

Pour calculer la densité prédictive en tout point à partir de ces N quantiles conditionnels, on a adopté les hypothèses utilisées lors du Challenge à savoir l'hypothèse que la densité conditionnelle soit uniforme entre deux valeurs successives de la cible et que les queues de distribution soit exponentielles [3]. Les différents estimateurs obtenus sont comparées sur le critère du NLPD sur le jeu de données test.
Nous avons tout d'abord constaté les mauvaises performances de l'estimateur Bayesien naïf utilisant l'ensemble des prédicteurs. Sur les trois jeux de données réels, le NLPD est en effet supérieur au NLPD pour la méthode dite de référence qui calcule à partir des données d'apprentissage l'estimateur empirique de la loi marginale $p(y)$. Lorsque l'hypothèse d'indépendance est trop forte, il est connu qu'elle dégrade fortement l'estimation des probabilités *a posteriori* (Frank et al., 1998). Ces mauvaises performances sont donc certainement dues à des corrélations importantes entre les prédicteurs.
Le tableau 2 indique le NLPD sur le jeu de données test pour les estimateurs MODL univarié et bivarié ainsi que pour la meilleure méthode du Challenge [4] et pour la méthode dite de référence. On observe tout d'abord que pour tous les jeux de données, les estimateurs MODL sont meilleurs que la méthode de référence, ce qui est loin d'être le cas pour toutes les méthodes soumises. On observe ensuite de bonnes performances des estimateurs MODL, notamment sur les jeux de données SO2 et Precip où les estimateurs univarié et bivarié se placent en tête. Les bonnes performances du prédicteur univarié montrent la qualité du partitionnement 2D obtenu malgré la manipulation exclusive des rangs et non des valeurs durant cette étape. D'autre part, l'estimateur bivarié est toujours meilleur que l'estimateur univarié. Cela indique la présence

[3] A cet effet, on a affecté une masse de probabilité $\epsilon = 1/N$ à chaque queue de distribution
[4] exceptée la soumission de l'organisateur

NLPD	Synthetic	SO2	Precip	Temp
Meilleure méthode soumise	0.386 (1er)	4.37 (2ème)	-0.279 (2ème)	0.034 (1er)
Bivarié MODL	$\times$	4.30 (1er)	-0.411 (1er)	0.25 (7ème)
Univarié MODL	1.02(5ème)	4.33 (1er)	-0.361 (1er)	0.285 (7ème)
Référence	1.23 (6ème)	4.5 (3ème)	-0.177 (4ème)	1.30 (9ème)

TAB. 2 – *Valeur du NLPD pour chacun des* 4 *jeux de données pour l'estimateur MODL univarié utilisant le meilleur prédicteur (Univarié MODL), l'estimateur MODL Bayesien naïf utilisant le meilleur couple (Bivarié MODL), la meilleure méthode soumise au Challenge et l'estimateur marginal de référence. Le classement de chaque méthode figure entre-parenthèses après chaque valeur de NLPD.*

d'informations supplémentaires et nous encourage à améliorer l'estimateur Bayésien naïf par le biais d'une sélection ou d'un moyennage.

5 Conclusion

Nous avons proposée une approche Bayésienne non paramétrique pour l'estimation de la loi conditionnelle du rang d'une cible numérique. Notre méthode se base sur un partitionnement 2D optimal de chaque couple (cible, prédicteur). Les effectifs de chaque partitionnement nous permettent d'obtenir des estimateurs univariés et un estimateur multivarié sous l'hypothèse Bayesienne naïve d'indépendance des prédicteurs. Une mise en œuvre de ces estimateurs sur des données proposés lors d'un récent challenge démontrent la qualité des partitionnements 2D. Les très bonnes performances du prédicteur univarié et du Bayesien naïf utilisant le meilleur couple de prédicteur nous encourage à travailler à l'amélioration du Bayesien naïf utilisant l'ensemble des prédicteurs.

Références

Boullé, M. (2006). MODL :a Bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Boullé, M. et C. Hue (2006). Optimal Bayesian 2d-discretization for variable ranking in regression. In *Ninth International Conference on Discovery Science (DS 2006)*.

Cawley, G., M. Haylock, et S. Dorling (2006). Predictive uncertainty in environmental modelling. In *2006 International Joint Conference on Neural Networks*, pp. 11096–11103.

Chaudhuri, P., M.-C. Huang, W.-Y. Loh, et R. Yao (1994). Piecewise-polynomial regression trees. *Statistica Sinica 4*.

Chaudhuri, P. et W.-Y. Loh (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernouilli 8*.

Chou, W. et Z. Ghahramani (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research 6*, 1019–1041.

Chu, W. et S. Keerthi (2005). New approaches to support vector ordinal regression. In *In ICML '05 : Proceedings of the 22nd international conference on Machine Learning*.

Crammer, K. et Y. Singer (2001). Pranking with ranking. In *Proceedings of the Fourteenth Annual Conference on Neural Information Processing Systems (NIPS)*.

D.J. Newman, S. Hettich, C. B. et C. Merz (1998). UCI repository of machine learning databases.

Fan, J., Q. Yao, et H. Tong (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika 83*, 189–196.

Frank, E., L. Trigg, G. Holmes, et I. Witten (1998). Naive Bayes for regression. Working Paper 98/15. Hamilton, NZ : Waikato University, Department of Computer Science.

Herbrich, R., T. Graepel, et K. Obermayer (2000). *Large margin rank boundaries for ordinal regression*, Chapter 7, pp. 115–132.

Hettmansperger, T. P. et J. W. McKean (1998). *Robust Nonparametric Statistical Methods*. Arnold, London.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monograph Series. Cambridge University Press.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B 42*, 109–127.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research 7*, 983–999.

Shashua, A. et A. Levin (2002). Ranking with large margin principles : two approaches. In *Proceedings of the Fiveteenth Annual Conference on Neural Information Processing Systems (NIPS)*.

Takeuchi, I., Q. Le, T. Sears, et S. A.J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research 7*, 1231–1264.

Summary

In some regression problems, we are more interested in predicting the rank of the output variable than its value. In this paper, we propose a non parametric Bayesian approach to estimate the complete distribution of the output rank conditionally to the input variables. For that, we first compute an optimal 2D partition which exhibit the predictive information of each input variable toward the output variable. We then show how we can build an univariate estimate and a naïve Bayesian estimate of the conditional distribution of the output ranks from the partition frequencies. These estimates are evaluated on four datasets proposed during the recent Predictive Uncertainty in Environmental Modelling Competition. If the naïve Bayesian estimate using all the input variables performs poorly, the univariate and the naïve Bayesian estimate using the best couple of input variables performs well despite their simplicity.